Measuring discrimination in algorithmic decision making

Zliobaite, Indre

2017

Zliobaite , I 2017 , ' Measuring discrimination in algorithmic decision making ' , Data Mining and Knowledge Discovery , vol. 31 , no. 4 , pp. 1060-1089 . https://doi.org/10.1007/s10618-017-0506-1

http://hdl.handle.net/10138/307578
https://doi.org/10.1007/s10618-017-0506-1

unspecified
acceptedVersion

Noname manuscript No.
(will be inserted by the editor)

Measuring discrimination in algorithmic decision making

Indre_ Zliobaite_

Received: [illegible] Revised [illegible] Revised 2nd [illegible]

Abstract *Society is increasingly relying on data-driven predictive models for automated decision making. This is not by design, but due to the nature and noisiness of observational data, such models may systematically disadvantage people belonging to certain categories or groups, instead of relying solely on individual merits. This may happen even if the computing process is fair and well-intentioned. Discrimination-aware data mining studies of how to make predictive models free from discrimination, when the historical data, on which they are built, may be biased, incomplete, or even contain past discriminatory decisions. Discrimination-aware data mining is an emerging research discipline, and there is no rm consensus yet of how to measure the performance of algorithms. The goal of this survey is to review various discrimination measures that have been used, analytically and computationally analyze their performance, and highlight implications of using one or another measure. We also describe measures from other disciplines, which have not been used for measuring discrimination, but potentially could be suitable for this purpose. This survey is primarily intended for researchers in data mining and machine learning as a step towards producing a unifying view of performance criteria when developing new algorithms for non-discriminatory predictive modeling. In addition, practitioners and policy makers could use this study when diagnosing potential discrimination by predictive models.*

Keywords *Discrimination-aware data mining* · *Fairness-aware machine learning* · *Accountability* · *Predictive modeling* · *Indirect discrimination*

I Zliobaite_
Dpt. of Computer Science, University of Helsinki, Finland;
Dpt. of Geosciences and Geography, University of Helsinki, Finland Email: indre.zliobaite@helsinki

Contents

1 Introduction

Nowadays, increasingly many decisions for people and about people are made using predictive models built on historical data, including credit scoring, insurance, personalized pricing and recommendations, automated CV screening of job applicants, profiling of potential suspects by the police, and many more cases. The penetration of data mining and machine learning technologies, as well as decisions informed by big data has raised public awareness that data-driven decision making may lead to discrimination against groups of people [1,10,18,21,23,33,44]. Such discrimination may often be unintentional and unexpected, assuming that algorithms must be inherently objective. Yet decision making by predictive models may discriminate against people, even if the computing process is fair and well-intentioned [5,14,15]. This is because most data mining methods are based upon assumptions that historical datasets are correct, and accurately represent population, which often appears to be far from reality.

Discrimination-aware data mining is an emerging discipline that studies how to prevent potential discrimination due to algorithms. It is assumed that non-discrimination regulations prescribe which personal characteristics are considered sensitive, or which groups of people are to be protected. The regulations are assumed to be dened externally, typically by national or international legislation. The research goal in discrimination-aware data mining is to translate those regulations mathematically into non-discrimination constraints, and develop predictive modeling algorithms that would be able to take into account those constraints, and at the same time be as accurate as possible. These constraints prescribe how much of dierences between groups can be considered explainable. In a broader perspective, research needs to be able to computationally explain the roots of such discrimination events before increasing public concerns lead to unnecessarily restrictive regulations against data mining.

In the last few years researchers have been developing discrimination-aware data mining algorithms using a variety of performance measures. Yet there is a lack of consensus of how to dene the fairness of predictive models, and how to measure their performance in terms of non-discrimination. Often research papers propose new ways to quantify discrimination, and new algorithms that would optimize that measure. The existing variety of evaluation approaches makes it dicult to compare results and assess progress in the discipline; furthermore, the variety of measures makes it dicult to recommend computational strategies to practitioners and policy makers.

The goal of this survey is to develop a unifying view towards discrimination measures in data mining and machine learning, and analyze the implications of optimizing one or another measure in predictive modeling. Therefore, it is essential to develop a coherent view early in the development of this research eld, in order to present task settings in a systematic way for follow up research, to enable systematic comparison of approaches, and to facilitate a discussion hopefully aimed at reaching a consensus among researchers in terms of the fundamentals of the discipline. For this purpose we review and categorize measures that have been used in data mining and machine learning, and also discuss measures from other disciplines, such as feature selection, which in principle could be used for measuring discrimination. We complement the review by experimental analysis of core measures.

Several surveys on dierent aspects of discrimination-aware data mining already exist and are complementary to this survey. A previous review [48] presents a multi-disciplinary context for discrimination-aware data mining. The review [48] focuses on approaches to solutions across dierent disciplines (law, economics, statistics, computer science), rather than analysis and comparison of measures. A yet earlier study [47] discusses a number of measures in relation to association rule discovery task, which in principle can be applied to any classication algorithm. This study discussed four measures that we current categorize under Absolute measures. A recent review [5] discusses the legal aspects of potential discrimination by machine learning, mainly focusing on American anti-discrimination laws in the context of employment, as well as discussing how big data and machine learning can lead to discrimination attributable to algorithmic effects regardless of jurisdiction. A classical handbook on measuring racial discrimination [8] focuses on surveying and collecting evidence for discrimination discovery. The book does not consider discrimination by algorithms, it only considers discrimination by human decision makers, and therefore presents inspiring ideas, but not solutions for measuring algorithmic discrimination, which is the focus of our survey. Interactions between human and algorithmic decision making is experimentally investigated in a recent study [6].

2 Background

The root of the word 'discrimination' is the Latin for distinguishing. While distinguishing is not undesirable as such, discrimination has a negative connotation when referring to adversary treatment of people based on belonging to some group rather than their individual merits. Initially associated with racism, nowadays discrimination may refer to a wide range of grounds, such as, race, ethnicity, gender, age, disability, sexual orientation, religion and more. Data mining is not aiming to decide what is the right or wrong reason for distinguishing, but considers sensitive characteristics to be externally decided by social philosophers, policy makers and society itself. The notion of sensitive characteristics can depend on the context and can change from case to case. The role of data mining is to understand generic principles and provide technical expertise on how to guarantee non-discrimination in algorithmic decision making.

2.1 Discrimination and law

Public attention to discrimination prevention is increasing, national and international anti-discrimination legislation are expanding the scope of protection against discrimination, and extending discrimination grounds. For instance, the EU is developing a unifying "Council Directive on implementing the principle of equal treatment between persons irrespective of religion or belief, disability, age or sexual orientation."

Adversary discrimination is undesired from the perspective of basic human rights, and in many areas of life non-discrimination is enforced by international and national legislation, to allow all individuals an equal prospect to access opportunities available in a society [24]. Enforcing non-discrimination is not only for the benefit of individuals. Considering individual merits rather than group characteristics is expected to benefit decision makers leading to more informed, and likely more accurate decisions.

From the regulatory perspective discrimination can be described by three main concepts: (1) what actions, (2) in which situations, and (3) towards whom are actions considered to be discriminatory. Actions are forms of discrimination, situations are areas of discrimination, and grounds of discrimination describe the characteristics of the people who may be discriminated against.

The EU legal framework for anti-discrimination and equal treatment is constituted by several directives, including the Race Equality Directive (2000/43/EC), the Employment Equality Directive (2007/78/EC), the Gender Recast Directive (2006/54/EC) and the Gender Goods and Services Directive (2006/113/EC) [62]. The main grounds for discrimination dened in European Council directives [17](2000/43/EC, 2000/78/EC) are: race and ethnic origin, disability, age, religion or belief, sexual orientation, gender and nationality. There is no general directive stating which attributes can and cannot be used for which types of decision-making [62]. Multiple discrimination occurs when a person is discriminated on a combination of several grounds. The main areas of discrimination are: access to employment, access to education, employment and working conditions, social protection and access to supply of goods and services.

Discriminatory actions may take dierent forms, the two main being known as direct discrimination and indirect discrimination. Direct discrimination occurs when a person is treated less favorably than another person would be treated in a comparable situation on protected grounds. For example, property owners not renting to a racial minority tenant. Indirect discrimination occurs where an apparently neutral provision, criterion or practice would put persons of a protected ground at a particular disadvantage compared with other persons. For example, the requirement to produce ID in the form of a driver's license for entering a club may discriminate against visually impaired people, who cannot have a driver's license. A related term statistical discrimination [2] is often used in economic modeling. It refers to inequality between demographic groups occurring even when economic agents are rational and non-prejudiced.

Data-driven decision making refers to using predictive models learned on historical data for decision support. Data-driven decision making is prone to indirect discrimination, since data mining and machine learning algorithms produce decision rules or decision models, which then may put persons of some groups at a disadvantage as compared to other groups. When decisions are made by human judgement, biased decisions may occur on a case-by-case basis. Rules produced by algorithms are applied to every case, and hence may discriminate more systematically and on a larger scale than human decision makers. Discrimination due to algorithms is sometimes referred to as digital discrimination [57].

The current non-discrimination legislation has been set up to guard against discrimination by human decision makers. The basic principles of the non-discrimination legislation generally apply to algorithmic decision making as well, the specics of algorithmic decision making are yet to be taken into national and international legislation. Ideally, algorithmic discrimination measures should be universal in a sense that they would not be tied to any specic legislation.

The current EU directives do not specify particular discrimination measures or tests to be used to judge whether there has been a discrimination. Rather, statistical measures of discrimination are used on case-by-case bases to establish prima facie *evidence, which then shifts the responsibility of proving discrimination from the person who is being discriminated against to the discriminating party.*

The general population, and even some data scientists may think that since data mining is based on data, models produced by data mining algorithms must be objective by nature. In reality models are as objective as the data on which they are built, and as long as the assumptions behind the models are perfectly matched in the data. In practice, assumptions are rarely perfectly matched. Historical data may be biased, incomplete, or record past discriminatory decisions that can easily be transferred to predictive models, and reinforced in new decision making [14]. Lately, awareness of policy makers and public attention to potential discrimination has been increasing [10,21,23,33,44], but there are many research questions which must be answered in order to fully understand in which circumstances algorithms do or do not become discriminatory, and how to prevent them being so by computational means.

2.2 Discrimination-aware data mining

Discrimination-aware data mining is a discipline at an intersection of computer science, law and the social sciences. It has two main research directions: discrimination discovery, *and* discrimination prevention. *Discrimination discovery aims at nding discriminatory patterns in data using data mining methods. A data mining approach for discrimination discovery typically extracts association and classication rules from data, and then evaluates those rules in terms of potential discrimination [28,39,40,47,49,53,54]. A more traditional statistical approach to discrimination discovery typically ts a regression model to the data including the protected characteristics (such as race or gender), and then analyzes the magnitude and statistical signicance of the regression slopes at the protected attributes (e.g. [22]). If those slopes appear to be signicant, then discrimination is agged. The majority of discrimination discovery approaches are based on nding correlations, whereas there is a growing body of research aimed at demonstrating causation [9,60], which is necessary for legal actions. Exploratory discrimination-aware data mining [6] is an emerging direction that aims to discover insights about new or changing forms of or grounds for discrimination. Discrimination-aware data mining relates to privacy-aware data mining (e.g. [29,52]) with a common understanding that securing privacy and non-discrimination come with a cost of information loss, and the objective is to minimize information loss while ensuring a desired level of privacy and fairness.*

Discrimination prevention algorithms have been developed to produce non-discriminatory predictive models with respect to externally given sensitive characteristics. The objective is to build a model or a set of decision rules that would obey non-discrimination constraints. Typically, such constraints directly relate to some selected discrimination measure. Algorithmic solutions for discrimination prevention fall into three categories: data preprocessing, model post-processing, and model regularization. Data preprocessing modifies historical data such that it no longer contains unexplained differences across the protected and the unprotected groups, and then uses standard learning algorithms with this modified data. Data preprocessing may modify the target variable [34,36,40] or modify input data [25,59] or both [28,29]. Model post-processing produces a standard model and then modifies this model to obey non-discrimination constraints, for instance, by changing the labels of some leaves in a decision tree [32,35] or removing selected rules from the set of discovered decision rules [30]. Model regularization forces non-discrimination constraints during the model learning process, for instance, by modifying the splitting criteria in decision tree learning [31,35,37]. Since the focus of this survey is on measuring discrimination, algorithmic solutions will be only briefly overviewed. An interested reader can find further details, for instance, in this edited book [19], this journal issue [43] or proceedings of specialized workshops [3,4,13].

Defining coherent discrimination measures is fundamental for both lines of research: discrimination discovery and discrimination prevention. Discrimination discovery requires some measure that can be used to judge whether there is any discrimination in data. Discrimination prevention requires some measure for use as an optimization criterion in order to sanitize predictive models. Direct discrimination by algorithms can be avoided by excluding the sensitive variable from decision making, but this unfortunately does not prevent the risk of indirect discrimination. In order to aid in establishing a basis for further research in the field, especially in algorithmic discrimination prevention, our main focus in this survey is to review indirect discrimination measures. While measuring direct discrimination is based on comparing individual to individual, measuring indirect discrimination is based on comparing group characteristics.

2.3 Definition of fairness for data mining

In the context of data mining and machine learning non-discrimination can be defined as follows: (1) people that are similar in terms of non-protected characteristics should receive similar predictions, and (2) differences in predictions across groups of people can only be as large as justified by their non-protected characteristics. To the best of our knowledge, in the data mining context these two conditions, expressed as Lipschitz condition and statistical parity, have been first formally discussed in [20].

The first condition is necessary but not sufficient for ensuring non-discrimination in decision making, because even though similar people are treated in a similar way, groups of similar people may be treated differently from other groups. The first condition relates to direct discrimination, which occurs when a person is treated less favorably than another would be treated in a comparable situation, and can be illustrated by the twin test. Suppose gender is the protected attribute, and there are two identical twins who share all the characteristics, but gender. The test is passed if both individuals receive identical predictions by the model.

The second condition ensures that there is no indirect discrimination, which occurs when apparently neutral provision, criteria or practice would put persons of a protected ground at a particular disadvantage compared with other persons. The so called *redlining* practice [32] exemplifies indirect discrimination. The term relates to past practices by banks to deny loans for residents of selected neighborhoods. Race was not formally used as a decision criterion, but it appeared that the excluded neighborhoods had much higher populations of non-white people than average. Thus, even though people of different races ("twins") from the same neighborhood were treated equally, the lowering of positive decision rates in the non-white-dominated neighborhoods affected the non-white population in a worse way. Therefore, different decision rates across groups of similar people can only be as large as explained by non-protected characteristics. The second part of the definition controls for balance across the groups.

More formally, let X be a set of variables describing non-protected characteristics of a person (a complete set of characteristics may not always be known or available, in such a case X denotes a set of available characteristics), S be a set of variables describing the protected characteristics, and $\hat{y}$ be the model output. A predictive model can be considered fair if: (1) the expected value for model output does not depend on the protected characteristics $E(\hat{y}|X, S) = E(\hat{y}|X)$ for all X and S, that is, there is no direct discrimination; and (2) if non-protected characteristics and protected characteristics are not independent, that is if $E(X|S) \neq E(X)$, then the expected value for model output within each group should be justified by some fairness model, that is $E(\hat{y}|X) = F(\hat{y}|X)$, where F is a fairness model. Defining and justifying F is not trivial, that is where a lot of ongoing effort in discrimination-aware data mining currently is concentrated.

Discrimination by predictive models can occur only when the target variable is polar, that is, some predictions outcomes are considered superior to

Siriella brevicaudata species group (Crustacea: Mysida: Mysidae) from the : H V W , Q G R 3 D F L ¿ F

Mikhail DANELIYA [1,*], W. Wayne PRICE [2] & Richard W. HEARD [3]

[1] Department of Biosciences, University of Helsinki, 00014 Helsinki, Finland.
[1] Taxonomicum, 01400 Vantaa, Finland.

urn:lsid:zoobank.org:author:04866F3A-61FA-4C37-8E6C-5D20F8ED6D17
[2] urn:lsid:zoobank.org:author:693DB9FE-3CF0-49A7-8CFA-D17560939FA0
[3] urn:lsid:zoobank.org:author:661DB91F-FBDE-4023-9515-F899504B430F

Abstract. The *Siriella brevicaudata* species group from the HVW ,QGR 3DFLILF DQG GHVLJQ